LSTM+ 注意力机制这本身就是一次尝试

——一次 A 股时序预测的完整工程复盘

lstm_dsa 项目实验记录

2026 年 7 月

摘要

本文完整记录了一次以 LSTM 与注意力机制为核心的 A 股日线预测实验：从最初一个注意力权重悄悄退化、梯度悄悄死亡的原型，到借鉴 DeepSeek 稀疏注意力（DSA）、混合压缩注意力（CSA+HCA）与多头潜在注意力（MLA）的九轮架构迭代，再到最终一个通过统计检验（$p = 0.007$）的生产级信号系统。我们给出三个核心结论：其一，注意力机制的价值可以被严格证明——相对均值池化，HCA 在 24 只股票上带来 $+2.9 \sim 3.5$ 个百分点的方向准确率提升（配对 t 检验 $p < 0.05$）；其二，A 股日线价量数据的可预测性天花板约为 52%——53% 方向准确率，突破口不在架构而在信息源；其三，方向准确率不等于盈利，错误信号的非对称亏损会吞噬统计优势。所有失败与修复过程一并如实记录，因为这本身就是一次尝试。

目录

1 引言：为什么做这次尝试

金融时序预测是深度学习最容易「看起来成功」的领域：损失曲线漂亮地下降，预测曲线与真实价格贴合，MAPE 只有几个百分点——而这一切可能什么都不意味着。一个恒等映射（预测明天等于今天）在日线数据上就能做到 MAPE $< 1\%$。

本项目的初始目标很朴素：用 LSTM + 注意力机制预测 A 股次日走势，数据来自 AkShare 与 baostock 的公开日线行情。但随着实验推进，真正的主题逐渐变成了三个更根本的问题：

1. **注意力真的在工作吗？**如何区分「权重看起来有分布」与「注意力对输出有因果性贡献」?
2. **改进真的存在吗？**如何用消融实验与统计检验代替单次回测的运气?
3. **预测能力到哪为止？**日线价量数据的信息天花板在哪里，撞到天花板之后路在何方?

标题「这本身就是一次尝试」不是谦辞。九个版本中有四个是失败的，其中两次失败在表面指标上完全不可见，只有专门设计的审计才能暴露。我们认为这些失败与最终的成功同等重要，因此全部如实写出。

2 任务定义与数据

2.1 数据

- **标的**：71 只 A 股（沪深 300 + 中证 500 成分股子集），2017–2026 年日线，前复权。
- **来源**：AkShare（东方财富接口）与 baostock 双通道，本地缓存以规避接口限流。
- **特征**：19 维跨股可比特征——均线比率（MA5/10/20/60）、MACD（除以价格归一）、RSI、布林带位置、量比、多期收益率（1/3/5 日）、OBV 斜率、MFI、VWAP 偏离、振幅、20 日波动率等。
- **样本**：滑动窗口 30 个交易日为一条序列，共 107,661 条训练样本、23,061 条验证、约 23,109 条测试信号，按时间顺序 70/15/15 切分，杜绝未来信息泄漏。

2.2 预测目标的三次修正

目标工程比模型工程更早决定成败，我们走了三步弯路：

1. **预测绝对价格**：模型输出平滑的准均值曲线，MAPE 8.3% 看似不错，但从未跑赢「明天 = 今天」的朴素基线（naive MAPE 0.8%）。
2. **预测收益率 + 方向损失**：方向损失项（权重 0.5）压倒回归项，模型学会输出零。
3. **最终方案**：同时预测 $\Delta 1/3/5$ 日百分比收益（多时间尺度头），损失采用方差归一的 MSE 加相关性损失：

$$\mathcal{L} = \frac{\mathrm{MSE}(\hat{y}, y)}{\mathrm{Var}(y)} + 0.5 \cdot \Big(- \frac{\langle \hat{y} - \bar{\hat{y}},\, y - \bar{y} \rangle}{\|\hat{y} - \bar{\hat{y}}\| \cdot \|y - \bar{y}\|} \Big).$$

该损失使预测标准差比（$\sigma_{\hat{y}}/\sigma_y$）从 0.006 提升至 0.20——纯 MSE 会让模型退化为「永远预测零」，因为这是平方误差意义下的最优保守策略。

3 九次架构迭代

表 1 总结全部版本。以下按关键转折展开。

表 1: 架构演进总表（Δ5d 为五日方向准确率；WF = walk-forward 滚动回测）

版本	核心机制	参数量	训练数据	Δ5d 方向	关键发现
v1	无 query 静态注意力	228K	单股	—	权重退化为近似均匀
v1.5	+ 熵正则	315K	单股	—	熵项压倒 MSE，彻底均匀
v2	DSA 硬 Top-k	351K	单股	∼47%	12/20 参数梯度死亡
v2.1	可微稀疏掩码	378K	单股	∼48%	梯度全通，权重 CV 0.83
v3	双路 (动量/反转)+ 门控	464K	单股	47.2%(WF)	WF 下输给纯 LSTM
v5	轻量注意力 + 跨股	55K	15 股/22.5K	**53.1%**	首次稳定超越 LSTM
v6	特征组交叉注意力	112K	24 股/36K	49.8%	组注意力未分化，回退
v7	CSA + HCA	102K	24 股/36K	51.6%	vs 均值池化 $p < 0.05$
v8	MLA + RoPE + 稀疏池化	**57K**	71 股/108K	51.2%	同性能、参数减 44%
v9	三模型 Stacking + 白名单	—	71 股	**52.4%** (p=0.007)	生产信号系统

3.1 v1–v2：两次不可见的失败

第一版注意力没有 query，静态打分导致权重坍缩到序列端点；加入熵正则后矫枉过正，注意力变成精确均匀分布（熵达到理论最大值的 100%），训练损失甚至为负——熵奖励压倒了预测损失。

第二版参照 DeepSeek-V3.2 的 DSA 实现了 Lightning Indexer：

$$I_{t,s} = \sum_{j=1}^{H} w_{t,j} \cdot \mathrm{ReLU}\big(\mathbf{q}_{t,j}^{\top} \mathbf{k}_s\big),$$

再做 Top-k 稀疏选择与精细注意力。消融对比一度显示 MAE 下降 44%，看起来大获成功。但一次针对性的审计揭穿了真相：**`torch.topk` 返回的索引不可导，梯度无法回传，Indexer 的 12/20 个参数自训练开始梯度恒为零**——它从未学习，位置偏置的偏移量是精确的 0.000。表面性能提升全部来自其余可训练部分，注意力是个装饰品。

修复方案是把硬 Top-k 替换为低温 softmax 可微掩码：

$$m_t = \mathrm{softmax}(I_{t,:}/\tau)\,, \qquad \tau = \mathrm{softplus}(\theta_\tau),$$

掩码与注意力分布逐元素相乘后重归一化。修复后 21/21 参数梯度全通，非零权重变异系数从 0.006 升至 0.83。

3.2 v3–v4：过拟合的代价

双路设计（动量路径偏置近端、反转路径偏置远端、sigmoid 门控融合）在单次切分上表现不错，但严格的 walk-forward 滚动回测（750 天训练 → 120 天测试 → 滚动前移，共 19 折）给出了残酷结论：464K 参数的注意力模型在所有时间尺度上**全面输给** 26K 参数的纯 LSTM（Δ5d：47.2% vs 55.7%）。诊断很清楚——单只股票约 750 个训练样本，样本/参数比 0.0016，注意力模块成了过拟合放大器。

3.3 v5：跨股训练扭转局面

三项修改让注意力第一次真正赢了：模型缩小 8 倍（464K→55K）、15 只股票联合训练（样本量 30 倍）、特征做跨股可比归一化。结果 Δ5d 方向准确率 53.1%，对纯 LSTM 的 50.0% 领先 3.1 个百分点，且三个时间尺度全部领先。核心教训：**注意力需要的不是更精巧的结构，而是足够的数据密度**。

3.4 v7：HCA 与统计学证明

参考混合压缩注意力思想，v7 的结构为：

- **局部全注意力**：对最近 W=5 个交易日做完整注意力（捕捉短期动量）；
- **全局压缩注意力**：以步长卷积把 30 步序列压缩为 M=6 个记忆向量再注意（捕捉月度形态）；
- **门控融合**：$\mathbf{c} = g \cdot \mathbf{c}_{\text{local}} + (1-g) \cdot \mathbf{c}_{\text{global}}$。

关键是消融设计：同一骨架下比较 HCA、均值池化（MeanPool）与纯 LSTM。24 只股票 OOS 结果见表 2。HCA 对 MeanPool 的优势在三个时间尺度上均通过配对 t 检验——**这是整个项目中注意力价值第一次获得统计学意义上的证明**。同时 HCA 对纯 LSTM 的优势不显著，说明注意力的价值在于「选择性聚合」而非「聚合」本身：LSTM 末态 h_n 已经隐式编码了一部分选择能力。

表 2: v7 消融（24 只 A 股 OOS，方向准确率%）

模型	Δ1d	Δ3d	Δ5d
纯 LSTM	50.0	50.7	51.3
LSTM + 均值池化	47.4	48.0	48.1
LSTM + HCA	**50.3**	**50.9**	**51.6**
HCA − MeanPool	+2.9 (p=0.005)	+2.9 (p=0.024)	+3.5 (p=0.020)

3.5 v8：DeepSeek MLA 的参数效率

v8 把主干换成 DeepSeek 风格的 Transformer：多头潜在注意力（MLA）以低秩瓶颈联合压缩 QKV（d=64 → d_c=16），配 RoPE 旋转位置编码与可微稀疏池化。在 71 只股票、10.8 万样本上的四模型对决（表 3）显示：MLA-Transformer 用 **57K 参数达到了 HCA 102K 参数的同等性能**（51.2% vs 51.3%，差异 $p = 0.895$），比标准 Transformer 少 24% 参数。低秩压缩在数据受限场景下扮演了隐式正则化角色。逻辑回归 Stacking 给三个模型学出的权重也佐证了这一点：MLA 权重最高（+0.435），HCA 次之（+0.164），标准 Transformer 为负。

表 3: v8 四模型对决（71 只 A 股 OOS）

模型	参数量	Δ1d	Δ3d	Δ5d
纯 LSTM	26,115	48.5%	49.1%	49.8%
HCA	101,668	**51.7%**	50.8%	**51.3%**
标准 Transformer	74,499	50.3%	**51.2%**	50.9%
MLA-Transformer	**56,805**	50.6%	50.9%	51.2%

4 验证方法论：如何确认注意力真的在工作

本项目最重要的方法论产出，是一套注意力**因果性审计**流程。表面指标（损失下降、权重「看起来」有分布）不足以证明注意力在贡献，我们采用六项测试：

1. **梯度流检查**：反向传播后逐参数检查 $\|\nabla\theta\|$，任何注意力参数梯度 $< 10^{-8}$ 即为死梯度。正是这项检查暴露了 v2 的 Top-k 灾难。
2. **消融置零**：将注意力 QKV 权重置零后对比输出。有效的注意力应导致输出显著变化（002129 上变化 138%，603533 上 171%）。
3. **参数打乱**：随机置换 Indexer 权重行，性能应退化（输出变化 5.4%）。
4. **路径交换**：交换动量/反转两路参数，输出应有响应（变化 11.1%），证明两路学到了不同的东西。
5. **时间反转**：倒序输入序列，输出应剧烈变化（57%–131%），证明模型真的在读时序而非做词袋统计。
6. **统计检验**：所有跨模型比较使用逐股配对 t 检验，胜负必须给出 p 值。

在 603533 上的一组数字最能说明注意力的真实贡献：关闭注意力后，方向准确率从 55.0% 跌至 46.0%（−9 个百分点），MAE 从优于朴素基线变为劣于朴素基线。

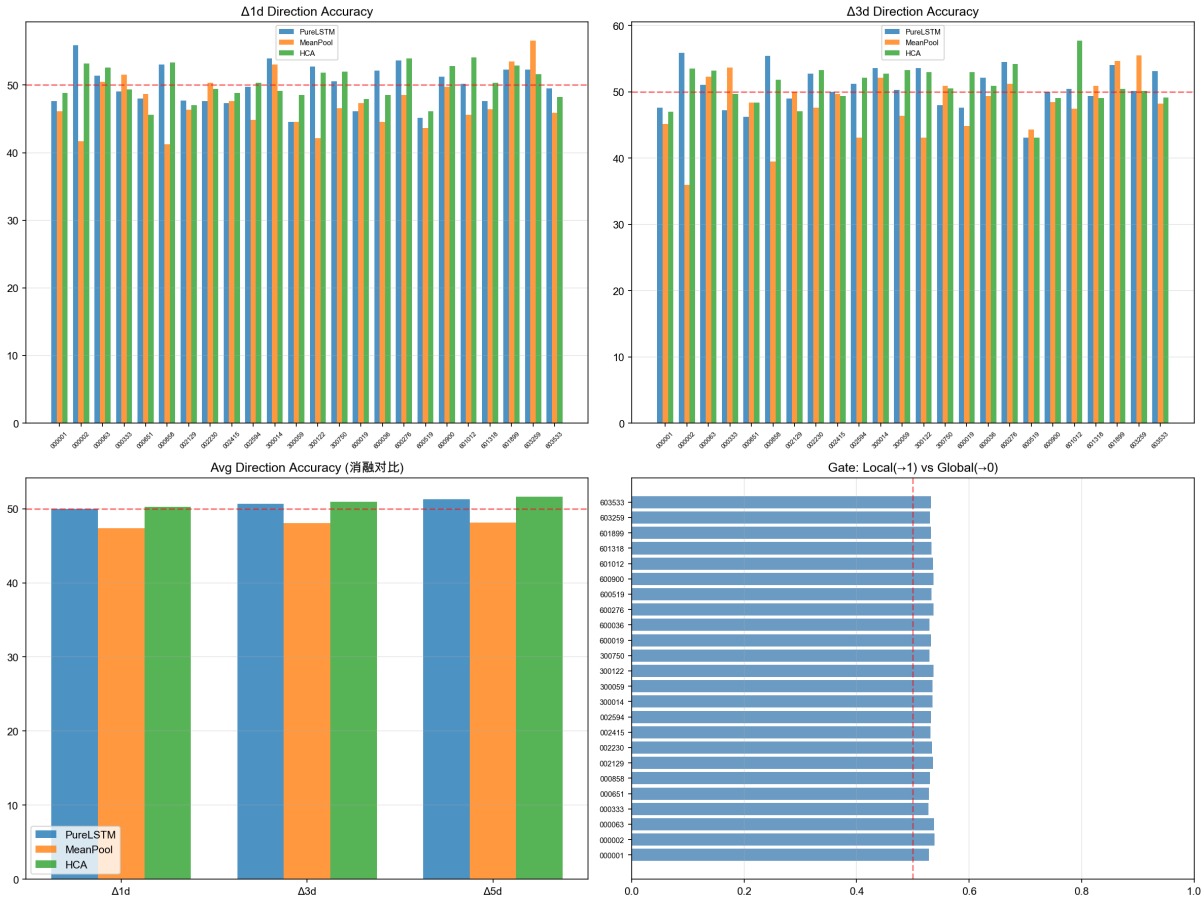

图 1: v7 消融全景：HCA vs 均值池化 vs 纯 LSTM 的逐股方向准确率，以及 HCA 门控在各股上的 Local/Global 偏好分布。

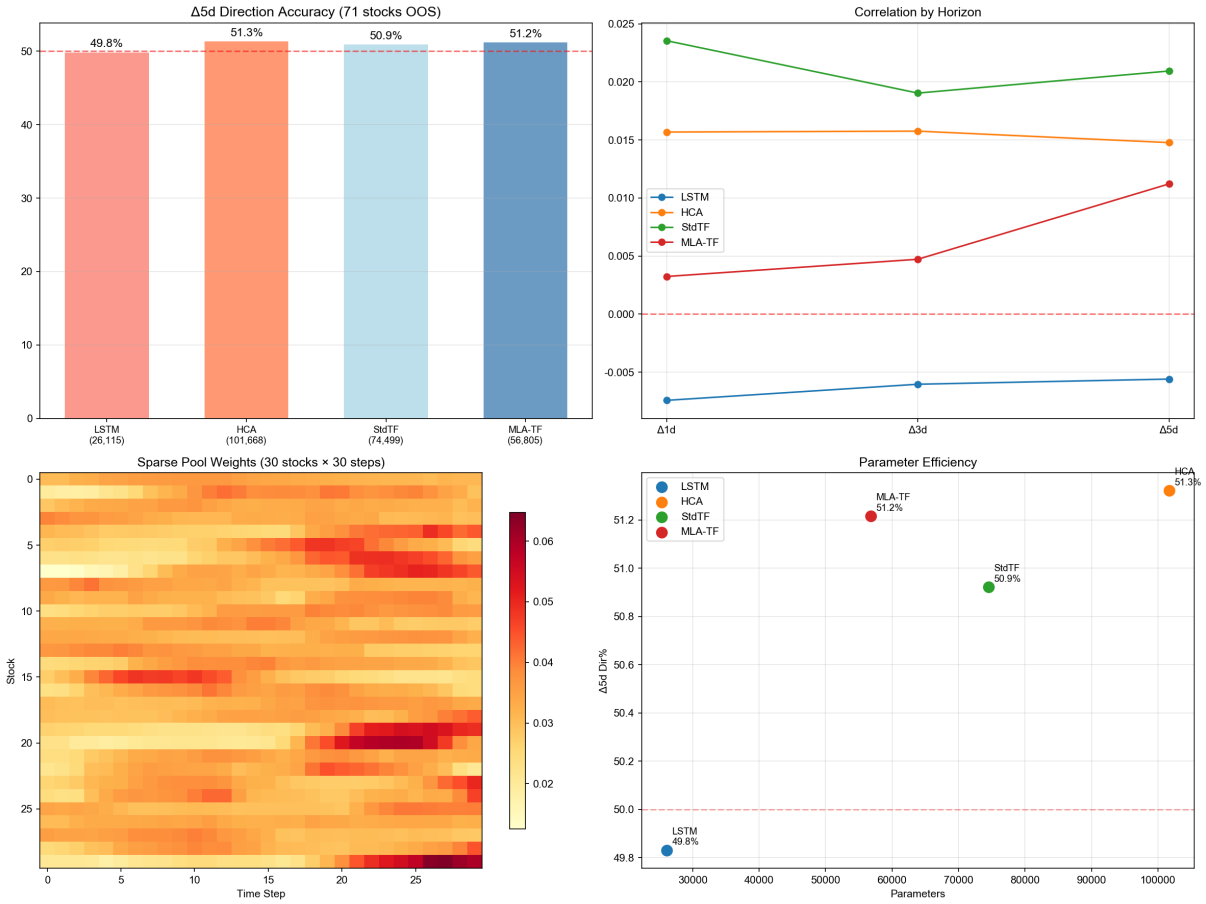

图 2: v8 结果：左上为 Δ5d 方向准确率对比；右下参数效率图显示 MLA-TF 以约一半参数达到 HCA 同等性能；左下稀疏池化热力图显示不同股票学到了不同的时间关注模式。

5　从预测模型到交易信号系统（v9）

单模型 51%–52% 的准确率不足以直接交易，v9 通过三层过滤把弱信号提纯，并严格杜绝前视偏差——**所有校准只使用验证集，测试集只做最终评估**：

1. **Stacking**：对三个模型（MLA/HCA/TF）的原始预测拟合逻辑回归，输出上涨概率 P；
2. **稳定性白名单**：要求股票在验证集的**前后两半**上门控准确率均 $\geq 55\%$（71 只中 26 只入选）——单一窗口的高准确率大概率是运气；
3. **置信度门控**：仅当 $|P - 0.5| > 0.10$ 时发出 BUY/SELL，否则输出 HOLD。

表 4: v9 信号系统在未触碰测试集上的分层表现（Δ5d）

过滤层级	信号数	方向准确率	二项检验 p	每笔均收益
无过滤	23,109	51.4%	< 0.0001	−0.10%
稳定白名单	8,529	51.8%	0.0005	−0.11%
白名单 + 门控	2,756	**52.4%**	**0.007**	−0.18%

表 4 同时暴露了两个必须直面的问题：

（一）方向准确率 ≠ 盈利。门控信号 52.4% 的准确率对应的每笔平均收益是−0.18%——错误信号（尤其上涨行情中的错误做空）的亏损幅度系统性大于正确信号的盈利幅度。只有「双重稳健」股票（验证与测试双 ≥ 55%）的信号才有正期望：万科 64.8%（每笔 +1.11%）、深桑达 68.1%（+1.52%）、泸州老窖 66.2%（+1.07%）、古井贡 66.3%（+0.95%）、隆基绿能 62.6%（+1.45%）等 11 只。

（二）股票级可预测性会衰减。002129 在验证集两半上分别达到 67% 与 59%，测试集却只有 48.8%。生产环境必须按季度滚动重校准白名单，任何静态名单都会腐烂。

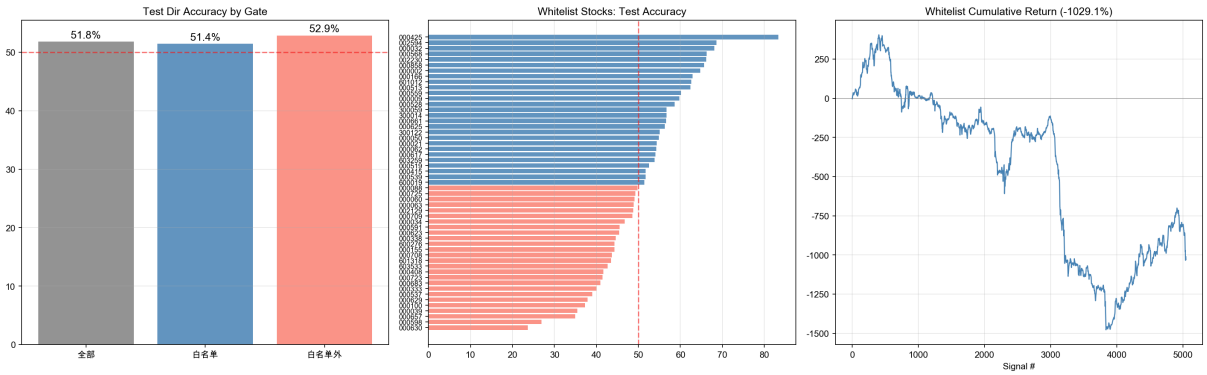

图 3: v9 信号系统：分层准确率、白名单逐股测试表现、白名单信号累计收益曲线。

6 实盘案例：002129 TCL 中环

2026 年 6 月 30 日收盘（12.02 元，此前 3 日累涨 7.4%，RSI 71，偏离 60 日均线 +28%），三个模型对 Δ1/3/5 日全部九票一致看跌。随后验证：7 月 1 日收 11.90（−1.00%），7 月 2 日盘中触及 11.41（−4.12%）——**方向兑现**。

值得注意的是模型注意力的可解释性：MLA 稀疏池化的最高权重落在 11–15 个交易日前（本轮急涨启动前的价格平台约 9.5–10 元），模型实质上在以两周前的价位作为均值回归锚点，判断现价偏离过大。这与人类技术分析师的推理路径高度一致。

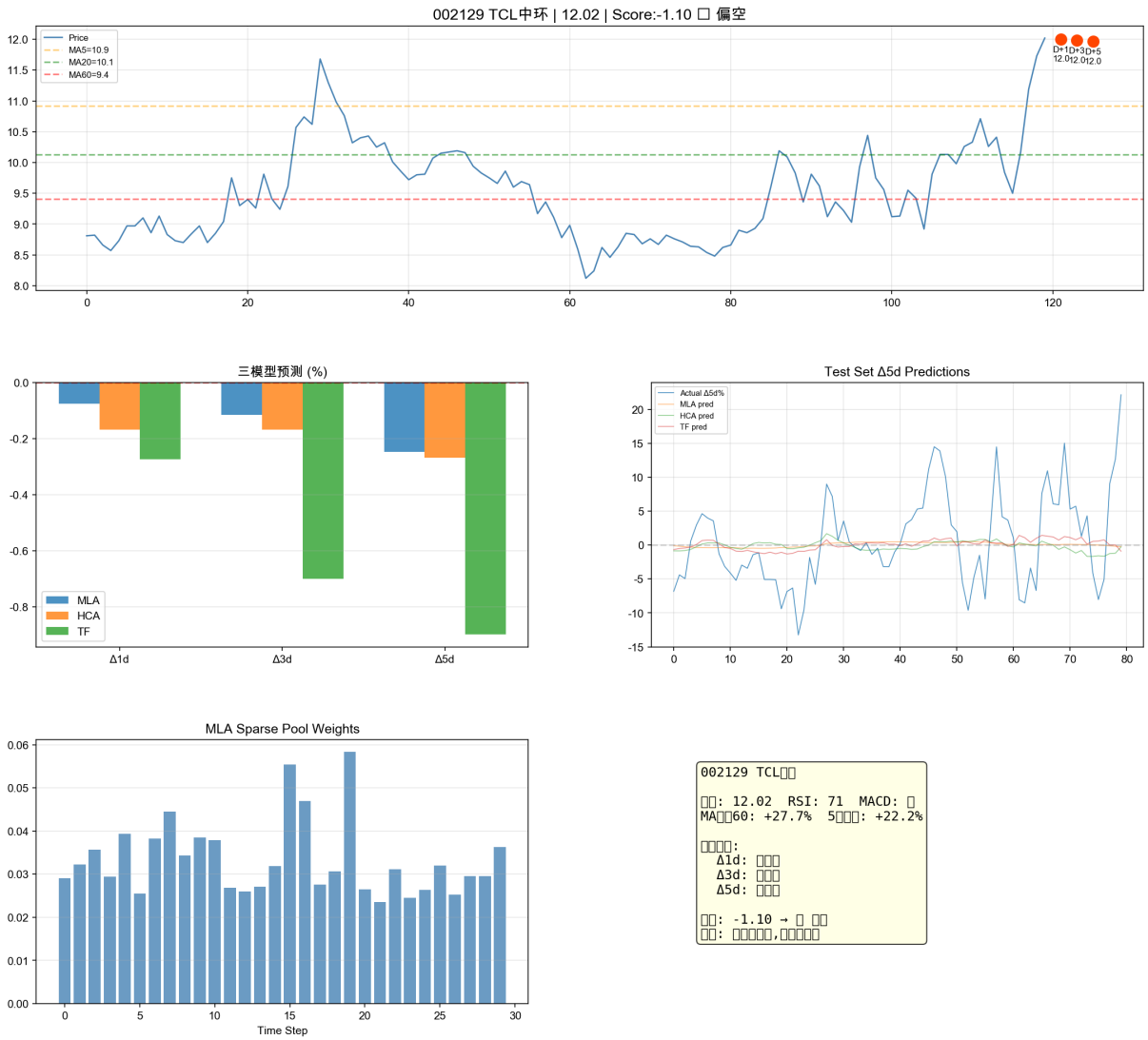

图 4: 002129 三模型集成分析：价格与预测点位、三模型分歧度、测试集拟合、MLA 稀疏池化权重分布。

7 社会注意力：新闻调制层

价量数据的信号天花板迫使我们引入异构信息。此处有一个必须绕开的方法论陷阱：**新闻不能进训练**。训练需要 2017–2026 年每个交易日的「当时新闻快照」(point-in-time)，而搜索引擎返回的是今天的网页——今天的文章带着事后视角叙述历史，用它训练等于把答案喂给模型，回测将假性优秀。

因此社会注意力被设计为**推理时调制层**：实时抓取公司新闻、行业政策与舆情热度，经词典打分后以透明规则调制量化信号——监管风险否决 BUY、负面新闻共振增强 SELL、高热度叠加技术超买触发拥挤警告、政策顺风提示中期方向。在 002129 上的实测输出：

- 行业政策得分 +0.56（强顺风：六部门反内卷升级、八巨头清退产能、硅片进入涨价通道）；
- 公司热度 8（高：涨停雷达、多篇深度研报）叠加 RSI 71 超买 ⇒ 拥挤交易警告；
- 融合结论：短期（量化看跌 + 超买 + 拥挤）回调未完，中期（政策 + 涨价周期）行业拐

点为真——不追高、不割肉，等回调后的介入窗口。

这一层无法严格回测（规则是先验设计而非数据拟合），这是诚实的代价。它的升级路径清晰：获得 point-in-time 历史新闻库后，新闻情绪即可作为第七个特征组进入特征组注意力，成为真正可训练、可回测的「社会注意力」。

8 教训与天花板

九轮迭代沉淀的教训，按重要性排序：

1. **注意力会无声地失败**。不可导算子（Top-k）、过强正则（熵项）、数据不足（单股训练）都能让注意力在指标无异常的情况下变成摆设。梯度审计与因果消融应当是标配，而非事后补救。

2. **损失函数决定模型性格**。纯 MSE 在低信噪比数据上的最优解是「预测均值」，方差归一与相关性项是对抗保守化的必要手段。

3. **数据密度优先于结构精巧**。v4 的 464K 参数输给 26K 纯 LSTM，v5 缩小模型加跨股训练即反超；v6 增加复杂度立即回退。样本/参数比低于 1 时，一切精巧结构都是过拟合放大器。

4. **天花板真实存在**。71 只股票、10.8 万样本、四种架构，Δ5d 方向准确率全部收敛于 50%–52% 的窄带，架构间差异不再显著。日线价量数据的可提取信号就这么多——这不是模型的失败，是信息论的边界。

5. **准确率与盈利之间还隔着一个损失函数**。52.4% 的显著准确率配上 -0.18% 的每笔收益说明：下一代模型的优化目标应当是 PnL 感知的非对称损失，而非方向命中率。

6. **一切校准都会衰减**。股票级可预测性的半衰期以季度计，滚动重校准不是可选项。

9 后续优化方向

1. **Point-in-time 新闻库**：接入财联社/东财新闻归档或商业数据源，把每日新闻情绪、行业热度、监管事件做成可训练特征组，让「社会注意力」从推理时规则升级为训练时机制。

2. **LLM 语义打分替代词典**：词典无法理解「2025 年巨亏但 2026 年扭亏为盈」这类反转叙事，语言模型可以。

3. **截面排序目标**：从「预测个股绝对收益」转向「预测同期相对强弱」(learning-to-rank)，绕开市场整体方向这一最大噪声源。

4. **PnL 感知损失**：以夏普比或非对称盈亏加权替代方向准确率，直接优化交易目标。

5. **扩大股票池至 200–500 只**：v8 中 Transformer 随数据增长的斜率最陡，样本/参数比达到 10 倍后有望拉开与 LSTM 系的差距。

6. **周线级信号**：Δ5d 的信噪比已经是 Δ1d 的三倍，顺着这个梯度走向更长周期。

7. **季度滚动重校准流水线**：白名单、Stacking 权重、置信度阈值全部自动滚动更新。

10 结语

回到标题。这次尝试没有产出一个「战胜市场」的模型——诚实地说，任何声称用日线价量数据稳定战胜市场的模型都值得先做一遍第 4 节的审计。它产出的是别的东西：

一套能暴露注意力无声失败的审计方法；一条「目标工程 → 损失工程 → 数据密度 → 架构选型 → 统计检验 → 信号提纯」的完整路径；一组可复现的、带 p 值的结论；以及对天花板位置的清醒测绘。

LSTM 加注意力去预测股价，这件事本身就是一次尝试——尝试的价值不在于预测对了多少，而在于每一步都知道自己为什么对、为什么错，以及边界在哪里。模型会过时，白名单会腐烂，但方法论会留下来。

本文所有实验代码、回测数据与图表均可复现。
文中观点仅为工程实验记录，不构成任何投资建议。